Publication of Statistical Linked Open Data in Japan

Yu Asano[1], Yusuke Takeyoshi[2], Junichi Matsuda[1], and Shoki Nishimura[3]

[1]Hitachi, Ltd.
[2]Oracle Corporation Japan
[3]National Statistics Center

Outline

1. Background and objective
2. Generation of statistical LOD
3. Publication and utilization of statistical LOD
4. Performance optimization for SPARQL query by Mr. Takeyoshi
5. Conclusion

1.1 Background

- From FY2008, “Portal Site of Official Statistics of Japan (e-Stat)” has been provided for publishing statistics of government agencies (Format: Excel)
- From FY2014, the API has been provided. (Format: XML, JSON)
- From FY2016, the statistical LOD site has been provided.

File download

FY2008-

Excel

500 statistics
1.1million tables

API

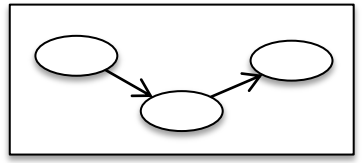
FY2014-

XML JSON

67 statistics
80,000 datasets

LOD

FY2016-



7 statistics
15 datasets

Reference
e-Stat (http://www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do)

1.2 Objective

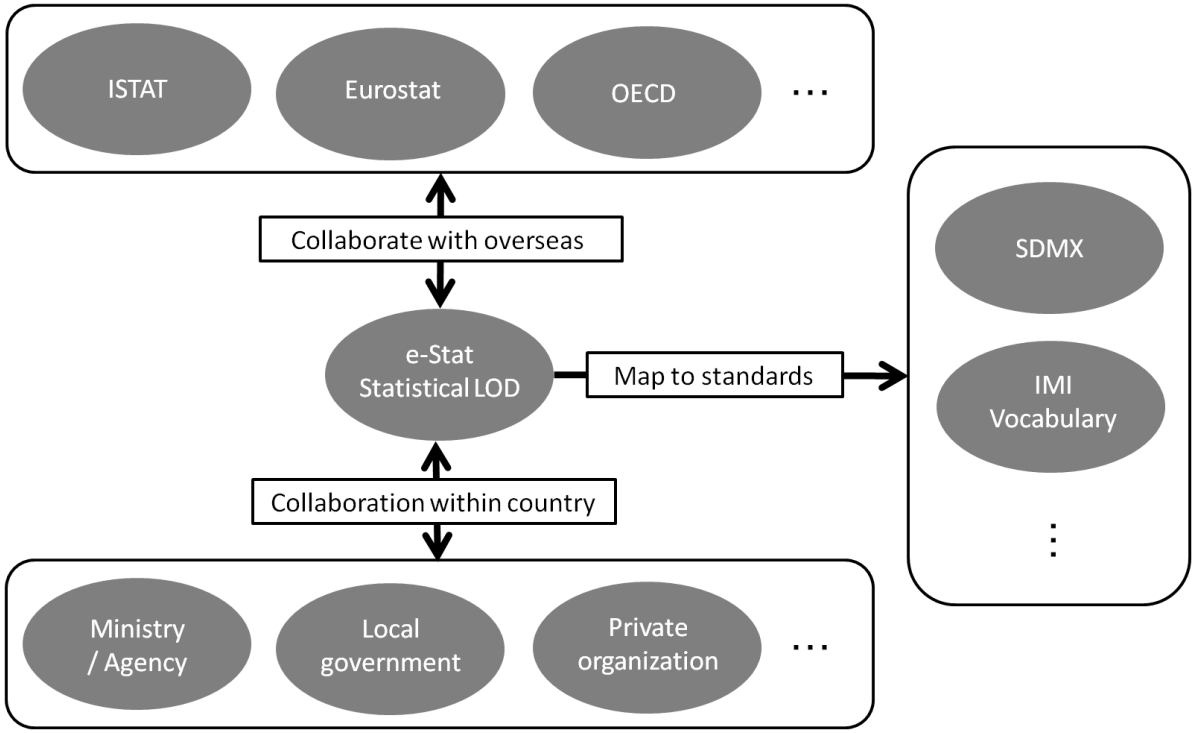
Objective: To promote domestic and international utilization of the statistics.
Proposal: To generate statistical LOD in consideration of the followings:

(A) Use of a unified structure: RDF Data Cube Model

(B) Use of a unified vocabulary:

- When a standard vocabulary exists: we use it.
- When a standard vocabulary does not exist: we define a new vocabulary

(C) Linking the defined vocabulary to external related vocabulary

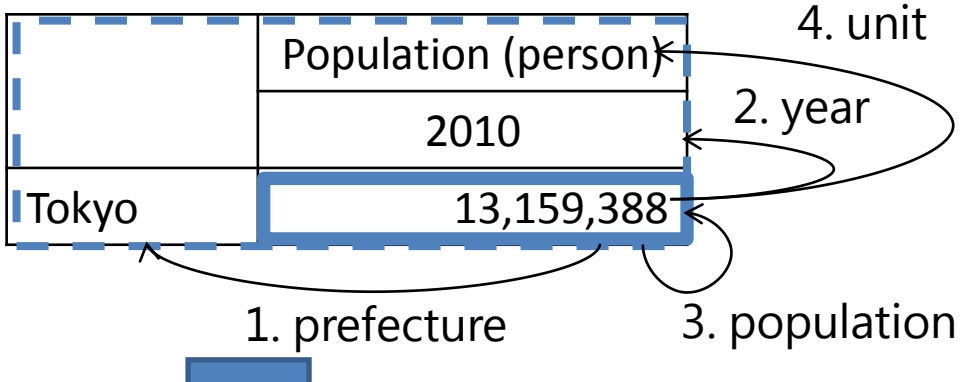


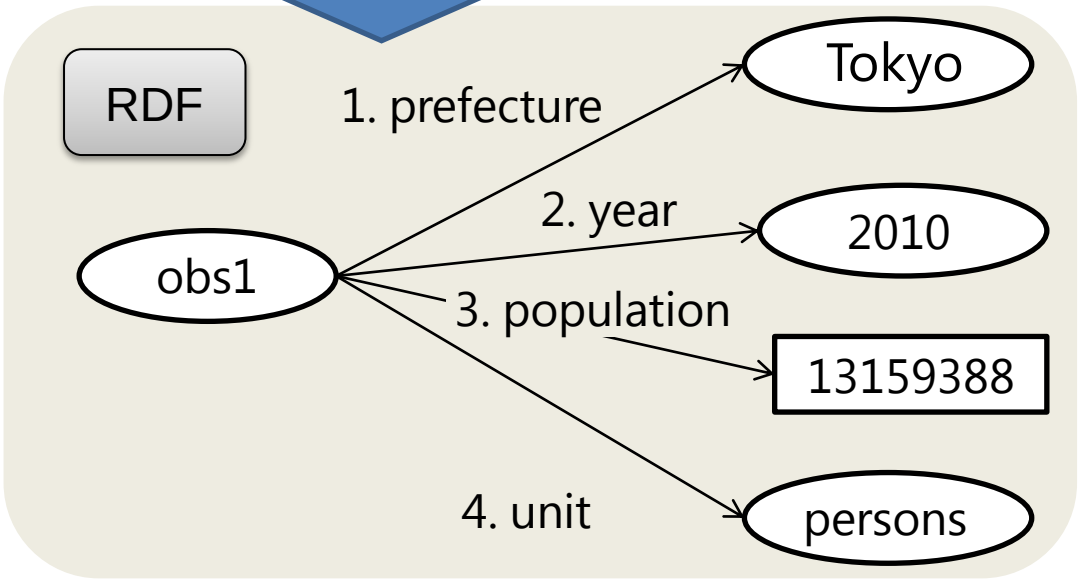
Statistical LOD by using RDF Data Cube Vocabulary

Statistical data are described per observation in one cell.
Each observation is described by using dimensions, measure, and attribute.

Example: Tokyo's population in 2010

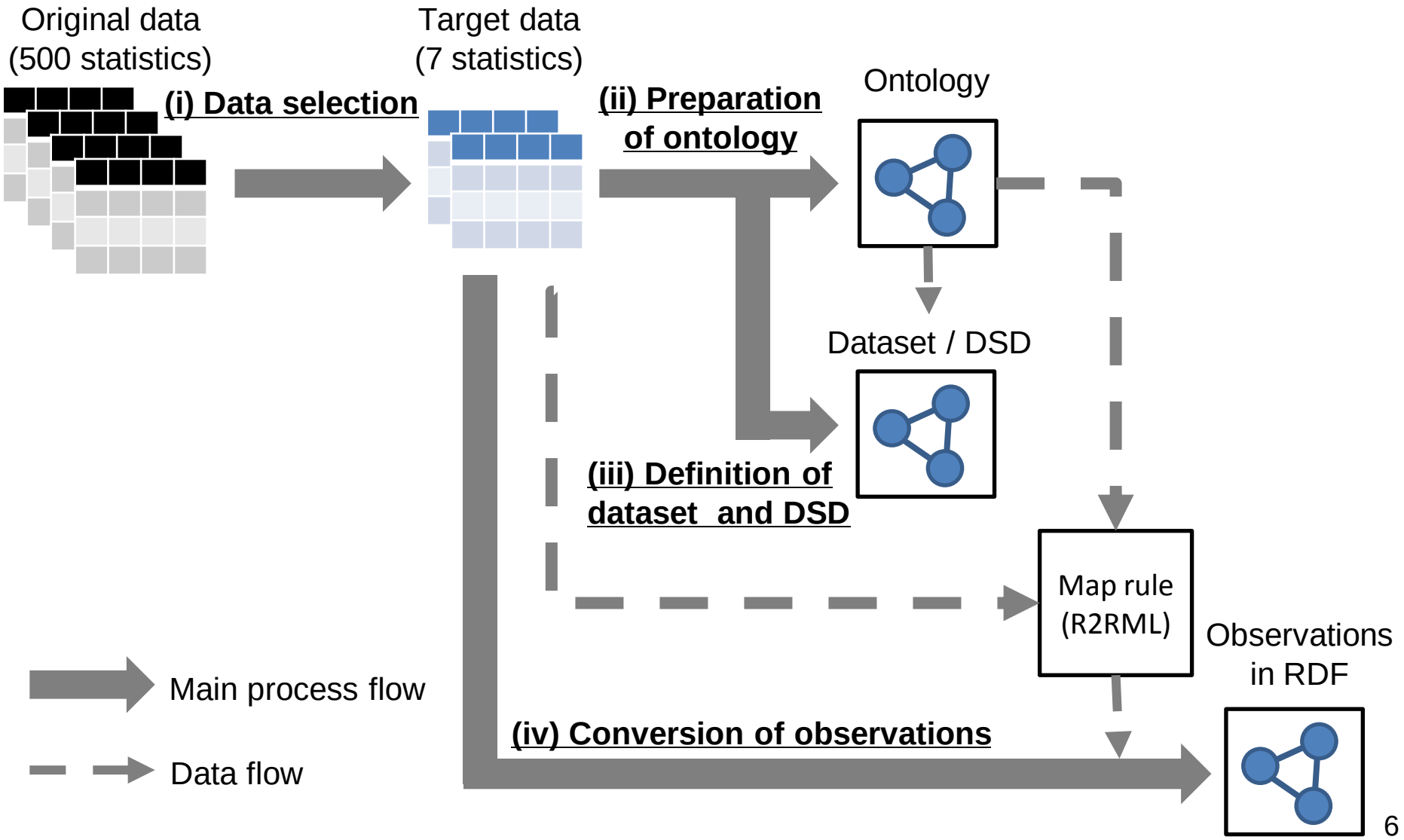
	Population (Persons)	
	2010	2009
Tokyo	13,159,388	12,988,797
Kanagawa	9,048,331	9,005,176
Chiba	6,216,289	6,183,743



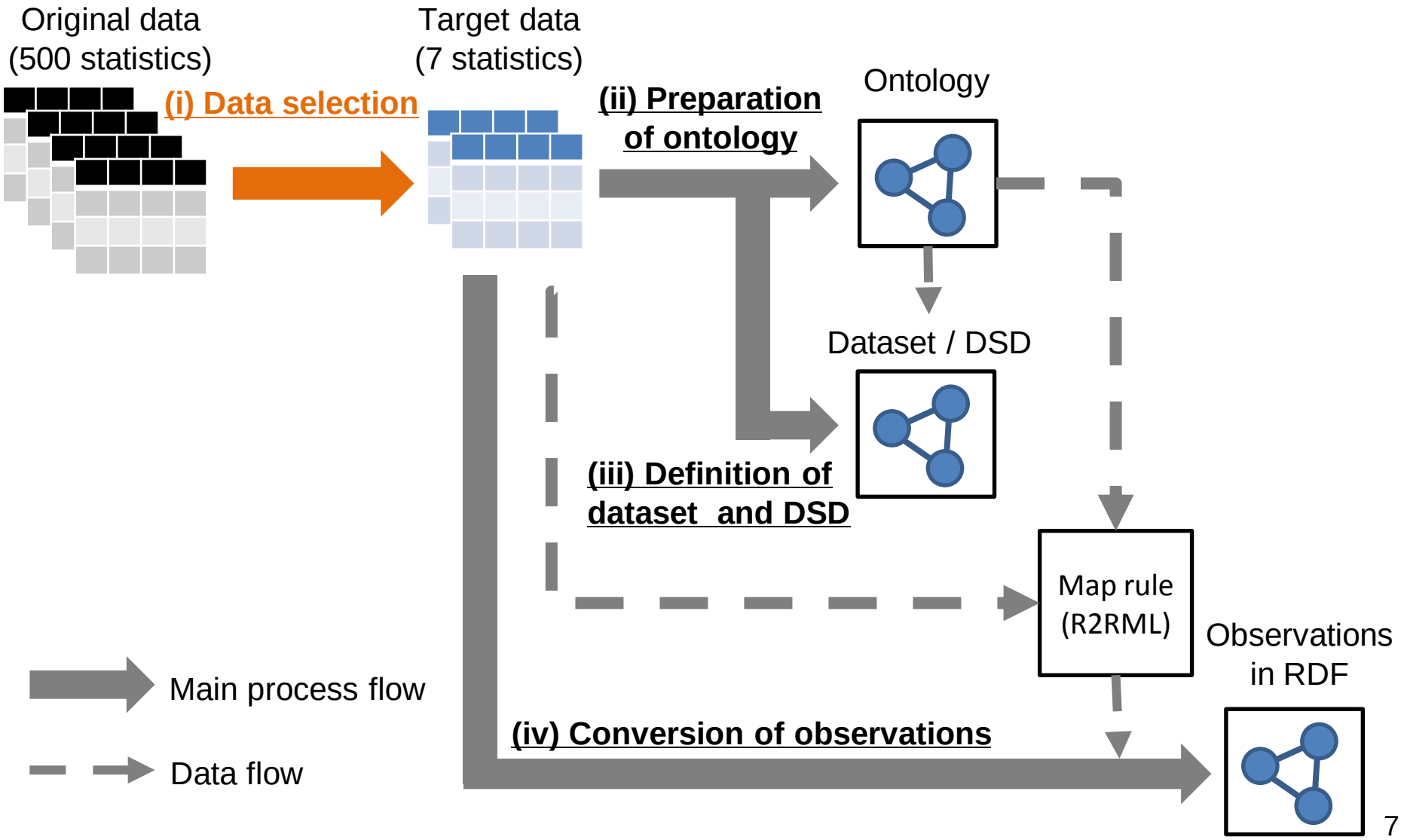


2. Generating method of statistical LOD

To convert a statistical data in RDB into LOD based on four stapes.

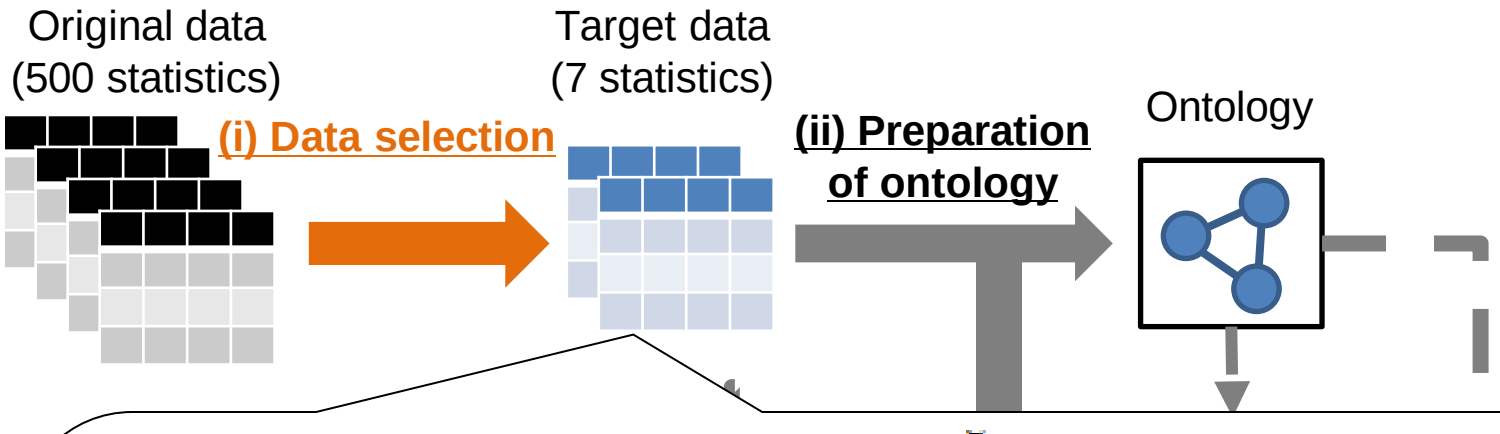


2. Generating method of statistical LOD



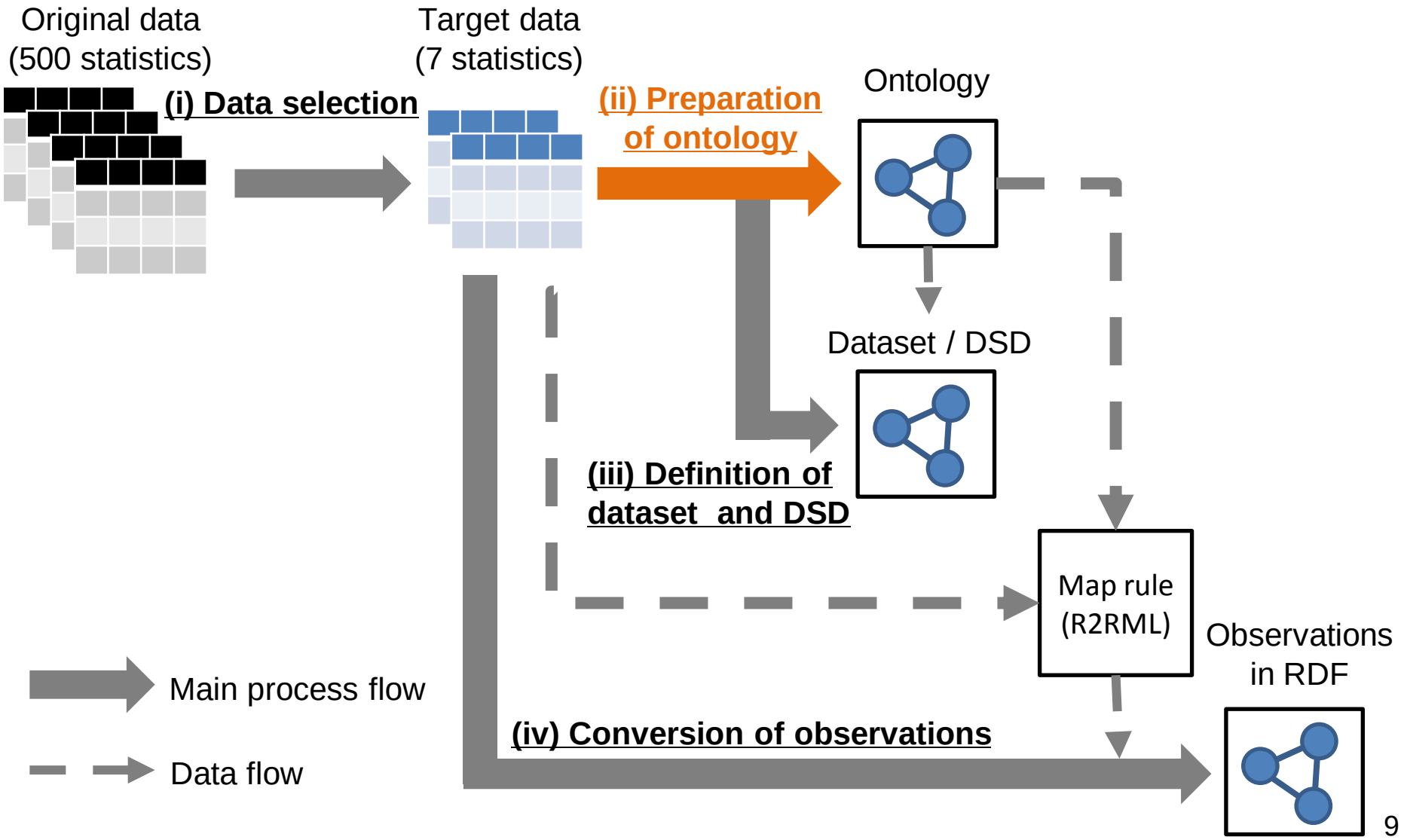
2. Generating method of statistical LOD

To select 7 statistics consisting of 15 datasets as LOD target in consideration of three points; (1) main statistics, (2) high needs, and (3) linking with external vocabularies.



Statistics	Tables	Measures
Population census (2010)	8	Population, Number of households etc.
Population estimation (2014)	1	Population (Population estimation)
Report on the internal migration in Japan (2014)	1	Number of people moving in from other municipalities, etc.
Economic census for business frame (2014)	2	Number of Establishments etc.
Labor force survey (Jan. 2012 -)	1	Labor force population, Employed population etc.
2010-Base Consumer price index (Jan. 2012 -)	1	Index
System of social and demographic statistics (2015)	1	Number of births, Number of elderly nursing facilities, etc.

2. Generating method of statistical LOD



2. Generating method of statistical LOD
(ii) Preparation of ontology

Step 1. List all necessary items for describing the target data
- Measures
- Dimensions
- Values of dimensions
- Attributes

Step 2. Check if there is the standard vocabulary or not
- Existence: Use it.
  e.g. Area code, IMI vocabulary (for Infrastructure for Multilayer Interoperability)
- Non-existence: Define it.

2. Generating method of statistical LOD
(ii) Preparation of ontology

Two considerations when defining items

(A) To give naming rules of the URI (Uniform Resource Identifier)
  (1) Uniqueness:
    - Use of own domain in URI (http://data.e-stat.go.jp/)
    - Use of name of statistics in URI
      (http://data.e-stat.go.jp/lod/ontoloty/populationCensus/dimension/2010/)
    - Use of “crossDomain” in URI when an item appears in multiple statistics
      (http://data.e-stat.go.jp/lod/ontoloty/crossDomain/dimension/2016/)
  (2) Manageability: Use of establishment year in URI for changeable items by year
      (http://data.e-stat.go.jp/lod/ontoloty/crossDomain/dimension/2016/)
  (3) Consistency: Use of unified naming rules for all items
  → Over 500 items were defined.

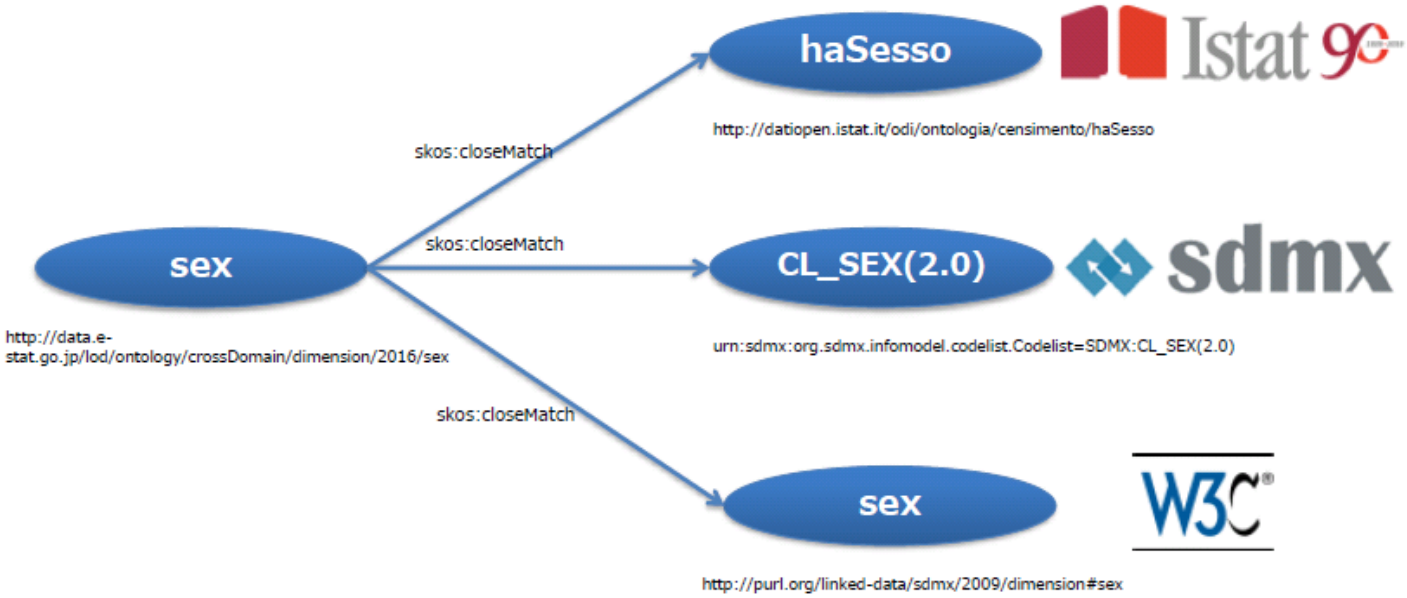
(B) To use SKOS (Simple Knowledge Organization System) to define a new vocabulary for linking the defined items to external related vocabularies

2. Generating method of statistical LOD
(ii) Preparation of ontology

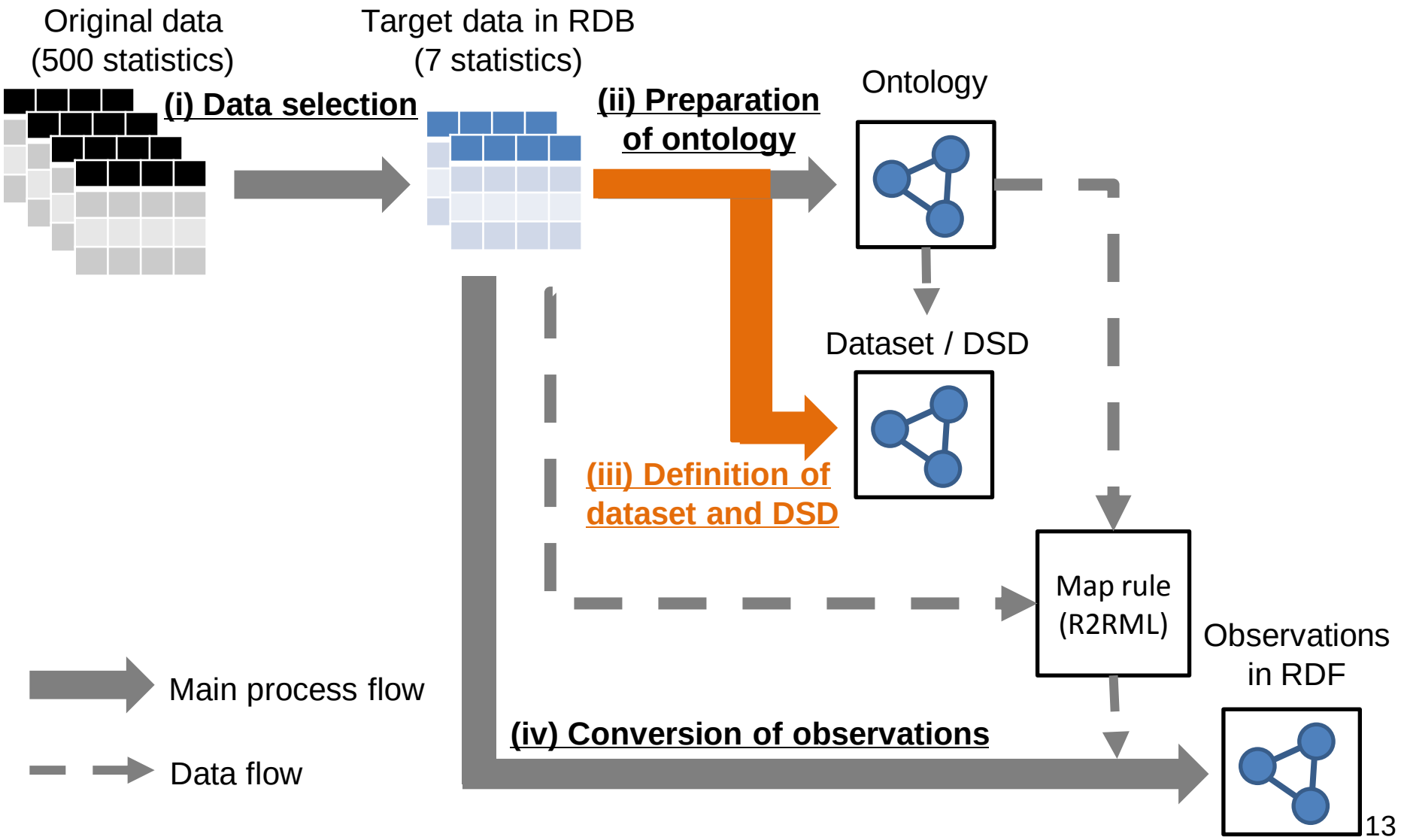
- We tried to link the defined items to the external related vocabularies from Italian statistics (Istat), Statistical Data and Metadata Exchange (SDMX), International Monetary Found (IMF), and SDMX (RDF encodings).
- 119 links are generated.

Table 2.2 Number of items linked to external vocabularies

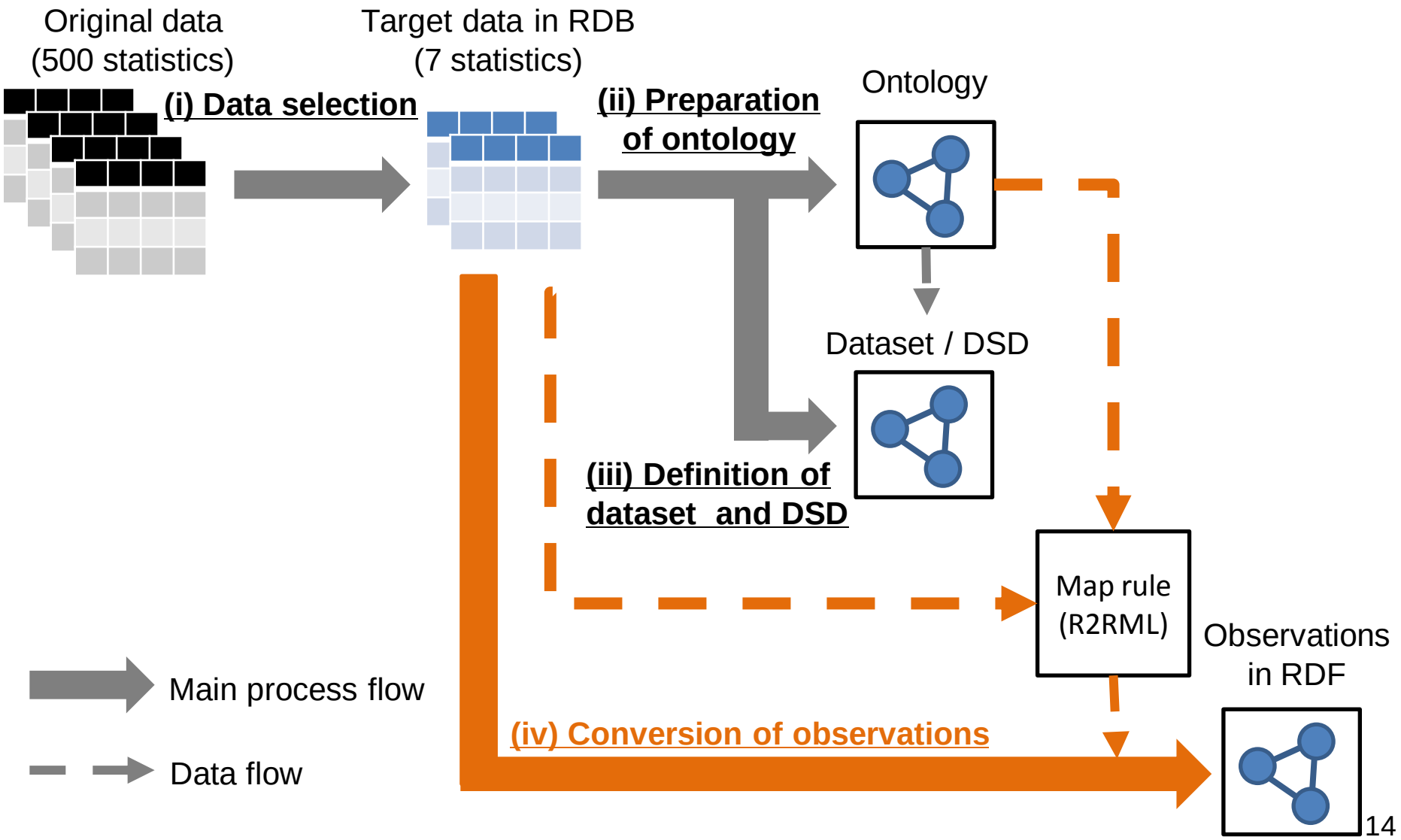
	Measures	Dimensions	Attributes	Values of dimensions	Values of attributes
Istat	3	5	0	22	0
SDMX	0	0	0	14	29
IMF	0	0	0	9	0
SDMX (RDF encodings)	0	7	2	11	17



2. Generating method of statistical LOD



2. Generating method of statistical LOD

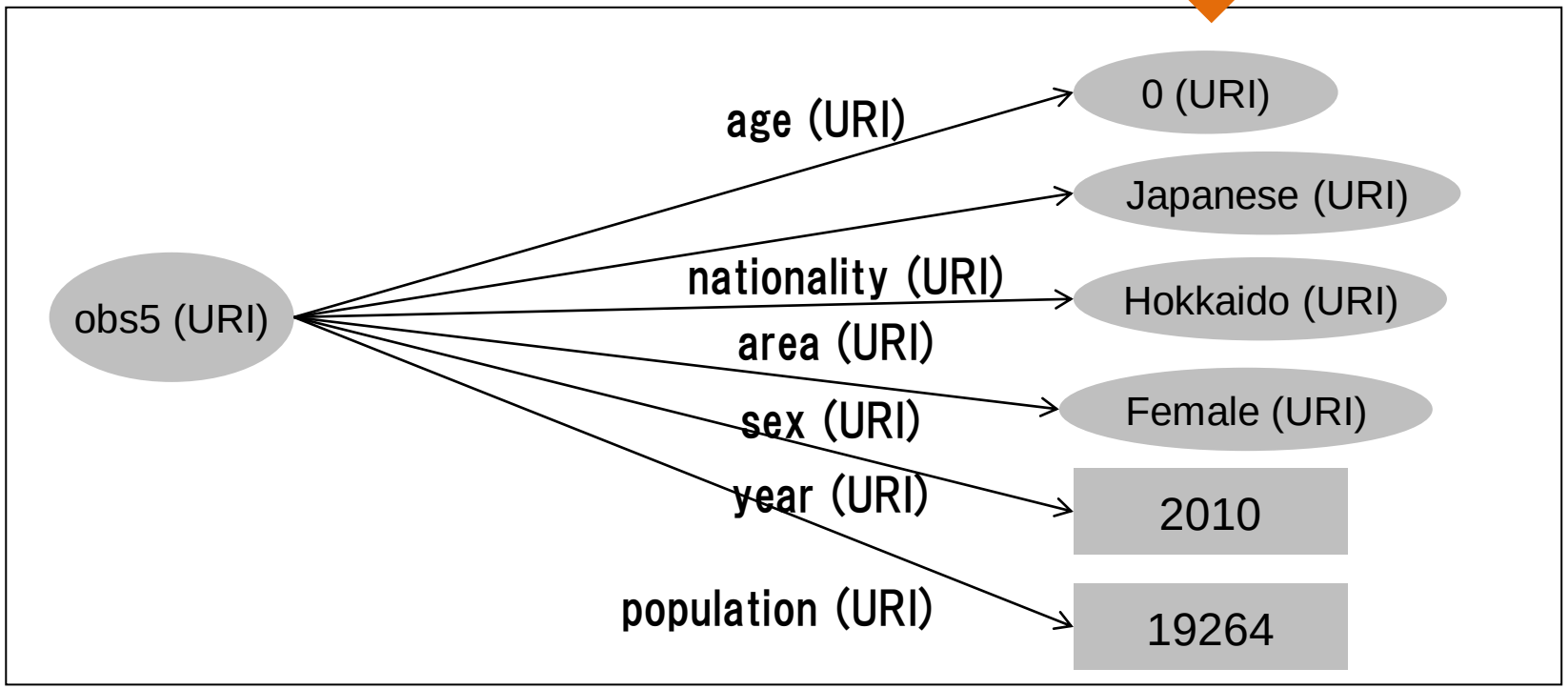


2. Generating method of statistical LOD
(iv) Conversion of observations

Example: The number of 0-year-old girls of Hokkaido in 2010

Table 2.1 Part of Population census 2010

			Total (Age)	0 year	1 year
Japanese (Nationality)	Hokkaido (Area)	Males	2,593,193	20,101	19,970
		Females	2,889,457	19,264	19,060



3. Publication and Utilization of statistical LOD

The statistical LOD site (http://data.e-stat.go.jp/lodw/) provides the followings.

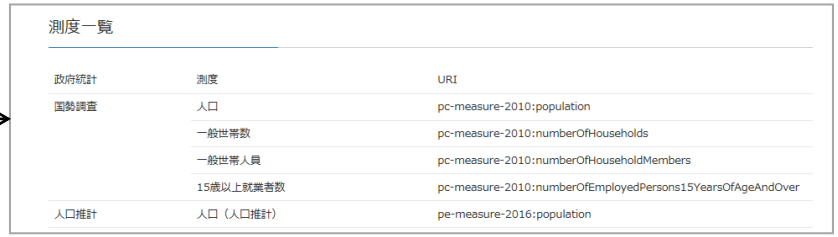
(1) SPARQL endpoint
    (7 statistics, 15 datasets, 300 million triples, 20 million observations)

(2) Definitions of items
- Datasets
- Data structures
- Dimensions
- Measures
- Attributes
- Codes

Top page

SPARQL endpoint

Definitions of items



Reference:
Linked Open Data (http://data.e-stat.go.jp/lodw/)

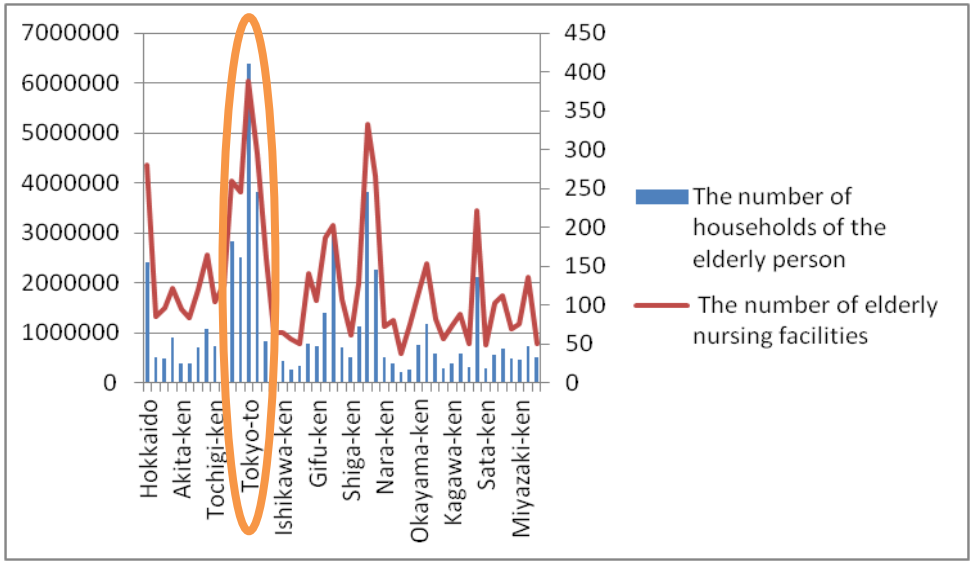
3. Publication and Utilization of statistical LOD

Example query for multiple statistics:

Q. Are the elderly nursing facilities enough or not in each area?

- The number of households of the elderly person by area (Population census)
- The number of elderly nursing facilities by area
  (System of social and demographic statistics)

```
SELECT  ?area ?numberOfHouseholds ?numberOfFacilities
WHERE { ?o1 pc-measure-2010:numberOfHouseholds ?numberOfHouseholds;
            pc-dimension-2010:typeOfHouseholdByPresenceOfAgedHouseholdMembers
                pc-code-2010:typeOfHouseholdByPresenceOfAgedHouseholdMembers-total;
            cd-dimension:standardAreaCode ?area_code.
        ?o2 ssds-measure-2016:J230121 ?numberOfFacilities;
            cd-dimension:standardAreaCode ?area_code.
        ?area_code sac:administrativeClass sac:Prefecture ;
        rdfs:label ?area.
        FILTER ( lang(?area)= "en" )
} ORDER BY ?area_code
```



Tokyo is less number of elderly nursing facilities for the number of the households of the elderly person than other areas.

Outline

1. Background and objective

2. Generation of statistical LOD

3. Publication and utilization of statistical LOD

4. Performance optimization for SPARQL query by Mr. Takeyoshi

5. Conclusion

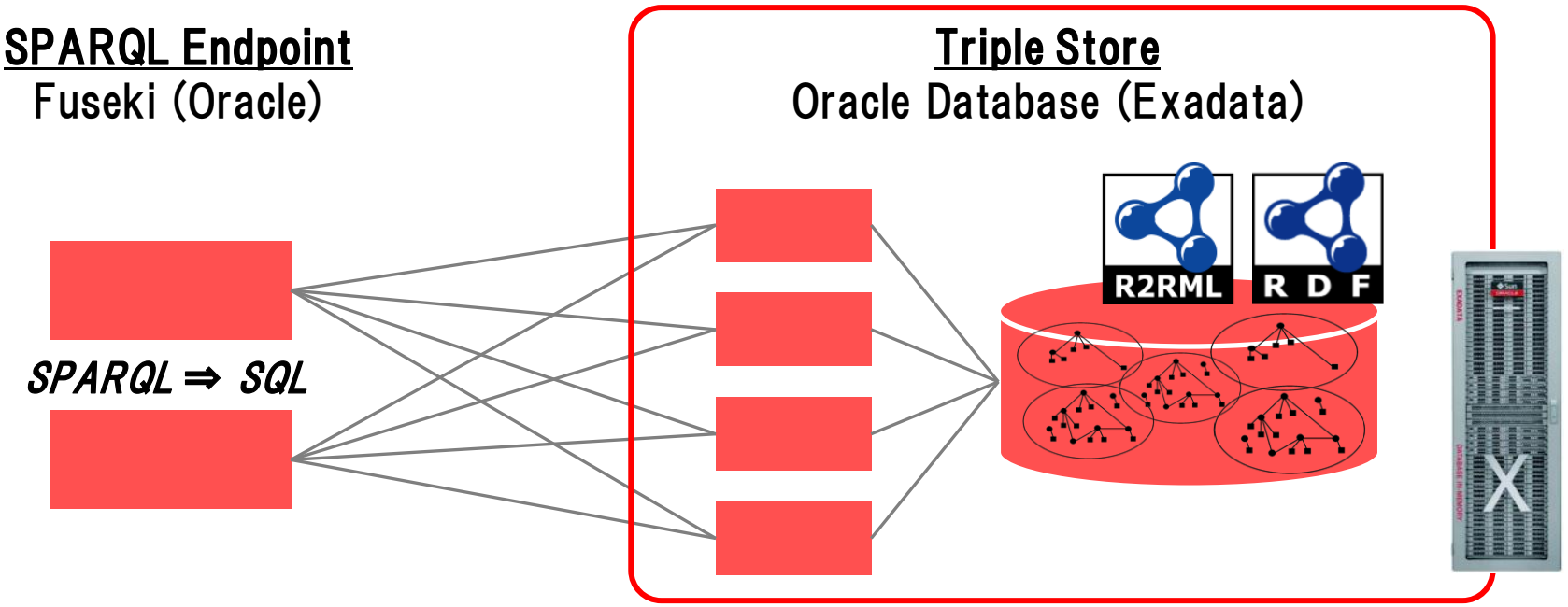
4. Performance optimization for SPARQL queries

System architecture and SPARQL processing flow

✓ The e-Stat LOD system is composed of Fuseki and Oracle Database on Oracle Exadata Database Machine.

1. SPARQL query is translated into semantically the same SQL at Fuseki.

2. The translated SQL query is processed in the database. The database only returns result sets to the Fuseki server.



4. Performance optimization for SPARQL queries

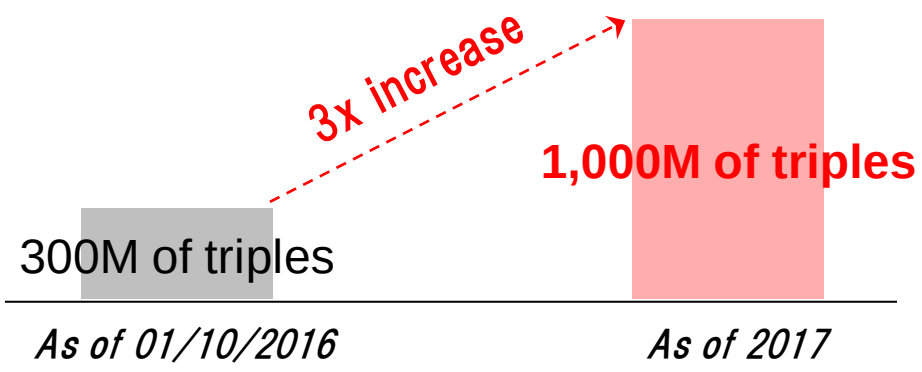
Performance concern in the e-Stat LOD

- ✓ The number of triples can easily increase when the original statistical tables have a large number of cells (observations), dimensions and measurements.

Table. Major published statistical datasets and number of triples

Statistical Dataset	Number of triples
Population Census	209,492,246
Economic Census for Business Frame	76,766,276

- ✓ While the total number of statistical tables published in the e-Stat LOD is expected to grow, an explosive increase in the number of triples could make SPARQL query response slower.



4. Performance optimization for SPARQL queries

Database design to improve SPARQL query performance

✓ We adopted the following three methods to achieve acceptable SPARQL performance even if there are more than billions of triples.

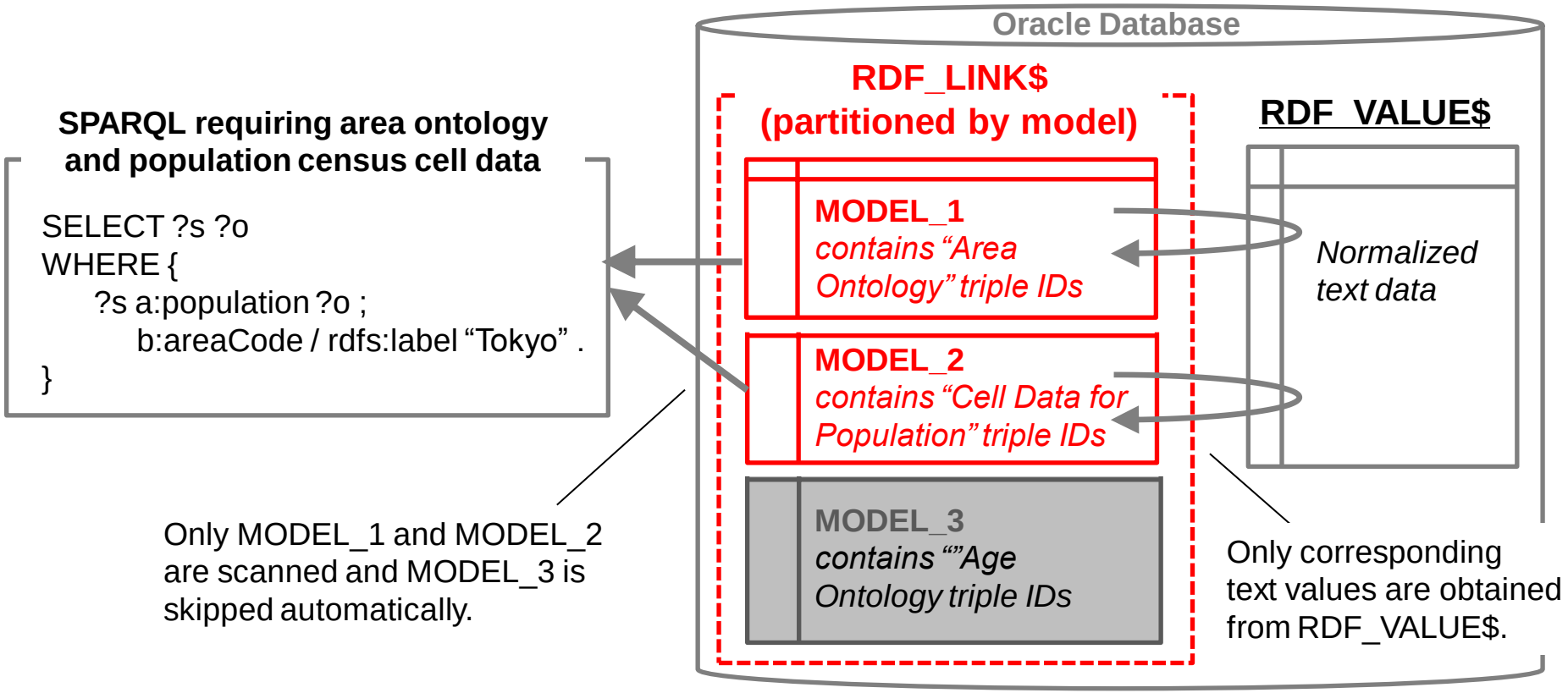
i. Storing each statistical dataset triples into different “models”
ii. In-database multi-process parallel processing for a single SPARQL query
iii. Daily automatic performance tuning for slow SPARQL queries

4. Performance optimization for SPARQL queries

Database design to improve SPARQL query performance

i. Storing each statistical dataset triples into different “models”

In the e-Stat LOD, semantically different datasets are stored in different “models”, which can limit the access range of SPARQL queries.

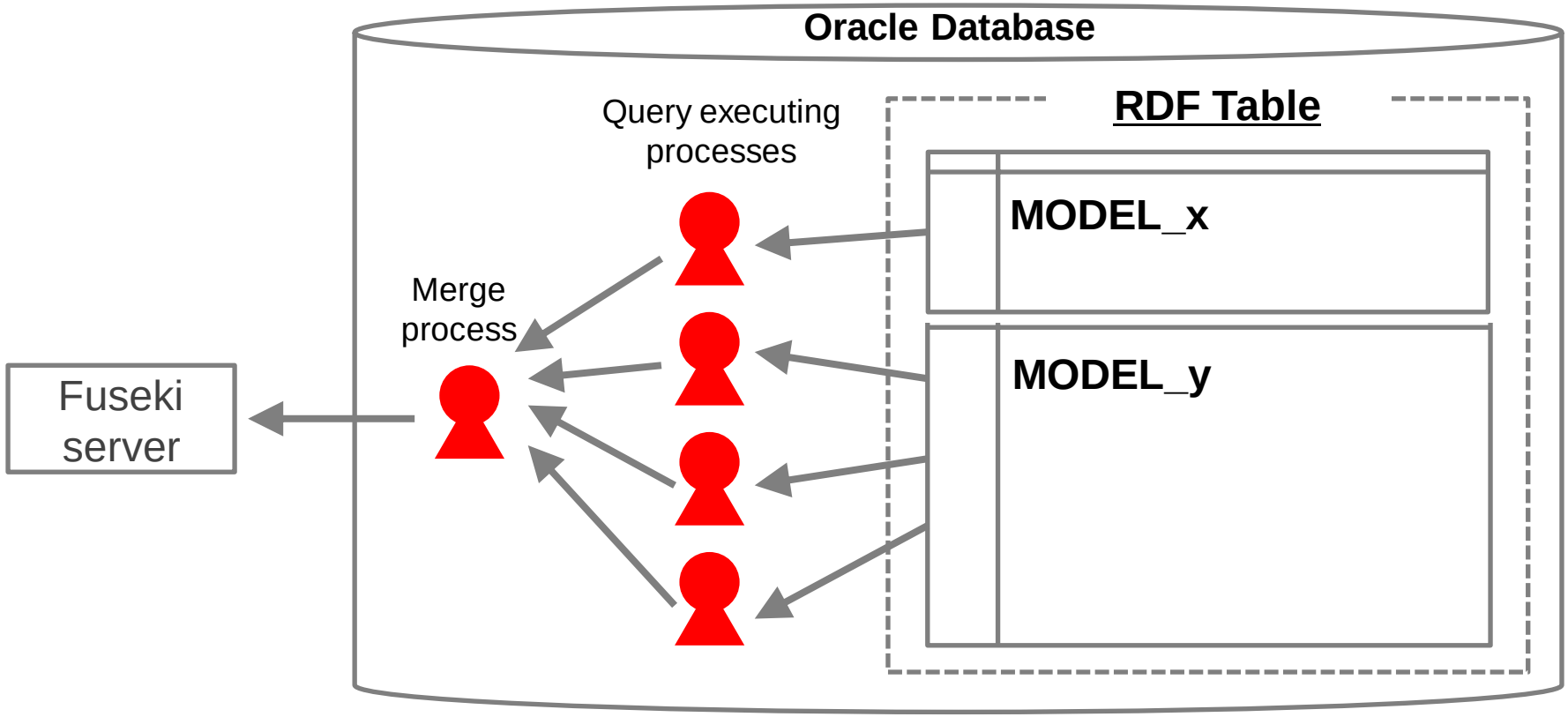


4. Performance optimization for SPARQL queries

Database design to improve SPARQL query performance

ii. In-database multi-process parallel processing for a single SPARQL query

A SPARQL query is automatically processed in parallel by multiple processes in the database when a query accesses a large portion of triples.

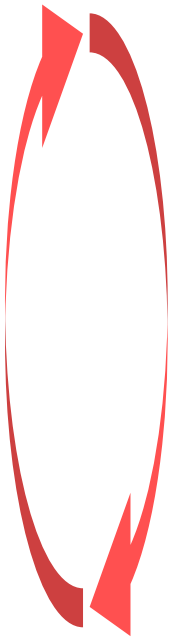


4. Performance optimization for SPARQL queries

Database design to improve SPARQL query performance

iii. Daily automatic performance tuning for slow SPARQL queries

Automatic SQL tuning job we implemented runs daily
to improve complex SPARQL response.



STEP1:
The performance tuning job picks up the top-10 slowest SQLs that are executed the day before.

STEP2
The job gathers various information for tuning, such as current condition of the target table/index data and the filter specified in the target SQL, and so on.

STEP3
Based on the information aggregated in STEP2, the job determines the optimal access methods for the target SQL, and automatically applies them to make the next execution of the same query faster.

Due to the repetition of the daily tuning job, the average SPARQL response is expected to get faster.

5. Conclusion

- Publication and utilization of statistical LOD in Japan
  - The LOD of 7 statistics was published
    - Use of a unified structure and vocabulary
    - Linking the defined vocabulary to external vocabularies

- Performance optimization for SPARQL queries
  - Storing each statistical dataset triples in different models
  - In-database multi-process parallel processing for a single SPARQL query
  - Daily automatic performance tuning for SPARQL queries slower than the criteria

- Future work
  - Expansion of LOD
  - Publication of usage guide and generation guide of LOD